university of groningen

University Medical Center Groningen

University of Groningen

Towards Empirical Evaluation of Affective Tactical NLG

van der Sluis, Ielka; Mellish, C.

Published in:
Empirical Methods in Natural Language Generation

DOI:
10.1007/978-3-642-15573-4_13

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Early version, also known as pre-print

Publication date:
2009

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
van der Sluis, I., & Mellish, C. (2009). Towards Empirical Evaluation of Affective Tactical NLG. In E. Krahmer, & M. Theune (Eds.), *Empirical Methods in Natural Language Generation* (pp. 146-153). (Lecture Notes in Computer Science; Vol. 5790). Springer. https://doi.org/10.1007/978-3-642-15573-4_13

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 05-05-2023

Towards Empirical Evaluation of Affective Tactical NLG

Ielka van der Sluis
Trinity College Dublin
Dublin
ielka.vandersluis@cs.tcd.ie

Chris Mellish
University of Aberdeen
Aberdeen
c.mellish@abdn.ac.uk

Abstract

One major aim of research in affective natural language generation is to be able to use language intelligently to induce effects on the emotions of the reader/ hearer. Although varying the *content* of generated language ("strategic" choices) might be expected to change the effect on emotions, it is not obvious that varying the *form* of the language ("tactical" choices) can do this. Indeed, previous experiments have been unable to show emotional effects of tactical variations. Building on what has been discovered in previous experiments, we present a new experiment which does demonstrate such effects. This represents an important step towards the empirical evaluation of affective NLG systems.

1 Introduction

This paper is about developing techniques for the empirical evaluation of affective natural language generation (NLG). Affective NLG has been defined as "NLG that relates to, arises from or deliberately influences emotions or other non-strictly rational aspects of the Hearer" (De Rosis and Grasso, 2000). It currently covers two main strands of work, the portrayal of non-rational aspects in an artificial speaker/writer (e.g. the work of Mairesse and Walker (2008) on projecting personality) and the use of NLG in ways sensitive to the non-rational aspects of the hearer/reader and calculated to achieve effects on these aspects (e.g. the work of De Rosis et al. (1999) on generating instructions in an emotionally charged situation and that of Moore et al. (2004) on producing appropriate tutorial feedback). Although there has been success in evaluating work of the first kind, it remains more problematic to evaluate whether work of the second type directly affects emotion or mood, or whether it influences task performance for other reasons.

Since the work of Thompson (1977), NLG tasks have been considered to divide mainly into those involving *strategy* ("deciding what to say") and *tactics* ("deciding how to say it"). It seems clear that one can affect a reader's emotion differently by making different strategic decisions about content (e.g. telling someone that they have passed an exam will make them happier than telling them that they have failed), but it is less clear that tactical alternations (e.g. involving ordering of material, choice of words or syntactic constructions) can have these kinds of effects. Unfortunately, the exact dividing line between strategy and tactics remains a matter of debate. For the purpose of this paper, we take "strategic" to cover matters of basic propositional content (the basic information to be communicated) and "tactical" to include most linguistic issues, including matters of emphasis and focus, inasmuch as they can be influenced by linguistic formulation. It is important to know whether tactical choices can influence emotions because to a large extent NLG research concentrates on tactical issues (partly because strategic NLG remains a rather domain-specific activity).

Some light on the effects of tactical variations in text is shed by work in Psychology, where there has been a great deal of work on the effects of the "framing" of a text (Moxey and Sanford, 2000; Teigen and Brun, 2003). Some of this has been industrially funded, as there are considerable applications, for instance, in advertising. The alternative texts considered differ in ways that NLG researchers would call tactical. For instance, a piece of meat could be described as "75% lean" or "25% fat", and arguably these are alternative truthful descriptions of the same situation. However, evaluation of this work has been primarily in terms of whether it affects people's *choices* or *evaluations*

of options available (Levin et al., 1998), or other aspects of task performance (O'Hara and Sternberg, 2001; Brown and Pinel, 2003; Cadinu et al., 2005). As far as we know it is unknown whether emotions can be affected in this way. There is therefore an open question about whether it is possible to detect the non-rational effects of different tactical decisions on readers. We believe that achieving this is important for the further scientific development of affective NLG.

In the rest of this paper, we discuss previous (unsuccessful) attempts to measure emotional effects of tactical decisions in texts (section 2), the particular linguistic choices we have focussed on, including a text validation experiment (section 3) and our choice of a method for measuring emotions (section 4). In section 5 we then present a new study which for the first time demonstrates significant differences in emotions evoked in readers associated with tactical textual variations. We then briefly reflect on this result in a concluding section.

2 Background for the Present Study

In (van der Sluis and Mellish, 2008) we described several experiments investigating different methods of measuring the effects of texts on emotions to demonstrate that tactical differences would lead to differences in effects. Our method was to present participants with texts about cancer-causing chemicals in foods or unexpected health-giving properties of drinking water and to attempt to measure the emotions invoked by different variations of these texts. However, we were unable to show statistically significant results of tactical variations. We mentioned the following possible explanations for this:

- We used methods where participants reported on their own emotions. However, it could be that (in this context) participants were unwilling or unable to report accurately.
- The self-reporting methods used were perhaps not fine grained enough to register the differences between the effects of similar texts.
- The texts themselves were perhaps too subtly different or not long enough to induce strong emotions.
- The participants were perhaps not involved enough in the task to get strong emotions.

We believe that of these, the final reason is the most compelling. The self-reporting methods used had been validated and used in multiple previous studies in Psychology, and so there was no reason to suggest that they would fundamentally fail in this new context. The granularity of the measurement methods can be improved relatively simply (see section 4 below). But it is very believable that the participants would fail to be really concerned by the texts in the experiments reported since the source was unclear, the message a general one not addressed to them individually and the topic (healthy and unhealthy food) one that occurs often enough in newspapers to fail to overcome natural boredom.

The main innovation of the experiment we describe below was in our method of seeking the emotional involvement of the participants. The texts that the participants read took the form of "feedback" on a (fake) IQ test that they undertook as part of the experiment. We selected university students as the participants, as they would likely be concerned about their intelligence, especially as compared to their peers. The texts appeared to be written individually for the participants and so sought to engage them directly.

3 Linguistic Choice and Framing

As in (van der Sluis and Mellish, 2008), the study we present here sought to evoke positive emotions to differing extents in a reader by tactical manipulations to "slant" the tasks positively to varying degrees. This section describes the text variations used and their validation.

3.1 Tactical Methods

The two texts produced for this experiment were written by hand, but used the following methods to give a more "positive slant" to a text. These are all methods that could be implemented straightforwardly in an NLG system[1]. In the following, the word "positive polarity" is used to refer to propositions giving good news to the reader or attributes which give good news to the reader if they have high values (such as the reader's intelligence). Similarly "negative polarity" refers to items that represent bad news, e.g. failing a test. For ethical reasons, negative polarity items did not arise in this experiment.

A. Sentence emphasis - include explicit emphasis in sentences expressing positive polarity propositions (e.g. exclamation marks and phrases such as "on top of this").

[1]Though the choice about *when* to apply them might not be so straightforward.

B. Choice of vague evaluative adjectives - when evaluating positive polarity attributes, choose vague evaluative adjectives that are more positive over ones that are less positive (e.g. "excellent", rather than "ok").

C. Choice of vague adverbs - provide explicit emphasis to positive polarity propositions by including vague adverbs expressing great extent (e.g. "significantly", rather than "to some extent" or no adverb).

D. Choice of verbs - for a positive polarity proposition, choose a verb that emphasises the great extent of the proposition (e.g. "outperformed", rather than "did better than").

E. Choice of realisation of rhetorical relations - when realising a concession/contrast relation between a positive polarity proposition and one that is negative or neutral, word it so that the positive polarity proposition is in the nucleus (more emphasised) position (e.g. say "although you did badly on X, you did well on Y" instead of "although you did well on Y, you did badly on X").

The idea is that an NLG system would employ methods of this kind in order to "slant" a message positively, rather that to present a message in a more neutral way. This might be done, for instance, to induce positive emotions in a reader who needs encouragement.

We claim that these choices can be viewed as tactical, i.e. that they are "allowable" alternative realisations of the same underlying content. For instance, we believe a teacher could use such methods in giving feedback to a student needing encouragement without fear of prosecution for misrepresenting the same truth that would be expressed without the use of these methods.

Whenever one words a proposition in different ways, it can be claimed that a (perhaps subtle) change of meaning is involved. However, in these cases we claim that it is the *writer's attitudes* that are being manipulated (and reflected in the text). We can therefore choose between these alternatives by varying the writer, not the underlying message. Our view is supported by a number of current accounts of the semantics of vague adjectives (though this is not an area without controversy). Many accounts of vagueness appeal to the idea that there is a norm which an adjective like "tall" implicitly refers to, and some of these argue both that the norm itself can be contextually determined and also that the amount by which the norm has to be exceeded has to be "significant" to a degree which is "relativized to some agent" (Kennedy, 2007). For instance, with the phrase "John is tall"

> "the property [...] attributed to John is not an intrinsic property, but rather a relational one. Moreover, it is not a property the possession of which depends only on the difference between Johns height and some norm, but also on whether that difference is a significant one. I take it that whether or not a difference is a significant difference does not depend only on its magnitude, but also on what our interests are" (Graff, 2000)

It is compatible with these accounts that different agents, with different interests and notions of what is noteworthy, can use vague adjectives in different ways[2].

Another reason for considering these methods as tactical is that in an NLG system, they would likely be implemented somewhere late in the "pipeline".

Probably the best way to check that we are using tactical alternations (according to our definition) is via some kind of text validation experiment with human participants. Section 3.3 below describes such an experiment, which provides strong support for this position.

3.2 Test Texts

For the experiment, we produced two feedback texts describing the same set of intelligence test results, one relatively neutral and one "positively slanted" using the above methods. In the experiment, they were given to participants in two groups, named "0" and "+" respectively. Each text consisted of 7 sentences, with a direct correspondance between the sentences of the two texts. Figure 1 presents the variations used in the feedback used in the experiment for group + (i.e. positively slanted) and group 0 (i.e. neutrally slanted). Note that the actual numbers are the same in both texts.

3.3 Text validation

A text validation study was conducted in which 15 colleagues participated. The participants were asked to comment on 12 sentence pairs, the 7 shown in Figure 1 and 5 additional filler pairs. The following analysis reports on our findings on the 7 sentence pairs shown in Figure 1 only.

In order that we could test our intuitions about the tactical nature of the linguistic alternations (discussed in section 3.1 above), the participants were presented with a scenario where there were two different teachers, Mary Jones and Gordon

[2]Though there are certainly *some* limits on the situations where a word like "tall" can be truthfully used to describe a height

+1: Your Baumgartner score of 7.38 is excellent!

01: Your Baumgartner score of 7.38 is ok.

+2: You did distinctively better than the average score obtained by other people in your age group.

02: You did somewhat better than the average score obtained by other people in your age group.

+3: Especially your scores on Imagination/Creativity and on Clarity of Thought were great and considerably higher than average.

03: Your scores on Imagination/Creativity and on Clarity of Thought were good and a little higher than average.

+4: A factor analyses of your Baumgartner score results in an overall excellent performance.

04: A factor analyses of your Baumgartner score results in an overall reasonable performance.

+5: Although, compared to your peers, you have only slightly higher Spatial Intelligence (7.5 vs 7.0) and Visual Intelligence (7.2 vs 6.8) scores, your Clarity of Thought Score is very much better (7.2 vs 6.3).

05: Compared to your peers, you have a somewhat better Clarity of Thought Score (7.2 vs 6.3), but you have only slightly higher Spatial Intelligence (7.5 vs 7.0) and Visual Intelligence (7.2 vs 6.8) scores.

+6: On top of this you also outperformed most people in your age group with your exceptional scores for Imagination and Creativity (7.9 vs 7.2) and Logical-Mathematical Intelligence (7.1 vs. 6.5).

06: You did better than most people in your age group with your scores for Imagination and Creativity (7.9 vs 7.2) and Logical-Mathematical Intelligence (7.1 vs. 6.5).

+7: There is a lot of variation in your age group, but your score is significantly higher than average.

07: Your score is higher than average, but there is a lot of variation in your age group.

Figure 1: Linguistic variation used in the IQ test feedback

Smith, both completely honest but with very different ideas about teaching (Mary believing that any pupil can succeed, given encouragement, but Gordon believing that most pupils are lazy and have overinflated ideas about their abilities). Given a positively slanted sentence (e.g. +7) from Mary and a corresponding more neutrally slanted one (e.g. 07) from Gordon, addressed to one or more pupils, participants were asked to indicate:

1. "Is it possible that Mary and Gordon might actually be (honestly) giving different feedback to the *same* pupil on the same task?"

2. "If the two pieces of feedback were given to the same pupil (for the same task) and the pupil's parents found out, do you think they would have grounds to make a complaint that one of the teachers is lying?"

The hypothesis was that (for the 7 pairs of sentences from Figure 1) in general participants would answer "yes" to question 1 and "no" to question 2. Indeed, for 6 pairs at least 14 out of the 15 participants answered as we had predicted. For the other pair (+4/04), 12 out of 15 agreed with both predictions. We see this as very strong evidence for our position (the participants gave different answers for the filler pairs, and so were not just producing these answers blindly).

No alterations were made to the two feedback texts on the basis of the text validation results.

4 Measuring Emotions

There are two broad ways of measuring the emotions of human subjects – physiological methods and self-reporting. Physiological methods unfortunately tend to have the problems of complex setup and calibration, which mean that it is hard to transport them between tasks or individuals. In addition, although emotional states are undoubtedly connected to physiological variables, it is not always clear what is being measured by these methods (cf. (Lazarus et al., 1980); (Cacioppo et al., 2000)).

Because of these problems, we have opted to investigate self-reporting methods, as validated and used widely in psychological experiments. Three well-established methods that are used frequently in the field of psychology are the Russel Affect Grid (Russell et al., 1989), the Self Assessment Manikin (SAM) (Lang, 1980) and the Positive and Negative Affect Scale (PANAS) (Watson et al., 1988). In our previous study (van der Sluis and Mellish, 2008), we had problems with participants understanding how to use the Russel Affect Grid and SAM and so now we opted to use a version of the PANAS test.

The PANAS test is a scale using affect terms that describe positive and negative feelings and emotions. Participants in the experiment read the terms and indicate to what extent they experience(d) the emotions indicated by each of them using a five point scale ranging from (1) very slightly/not at all, (2) a little, (3) moderately, (4) quite a bit to (5) extremely. A total score for positive affect is calculated by simply adding the scores for the positive terms, and similarly for negative affect.

As before, we used a simplified version of the PANAS scale in order not to overburden the participants with questions and to avoid bored answering. In this test, which has been fully validated (Mackinnon et al., 1999), participants have to rate only 10 instead of 20 terms: 5 for positive af-

fect (i.e. alert, determined, enthusiastic, excited, inspired) and 5 for negative affect (i.e. afraid, scared, nervous, upset, distressed).

Our use of the simplified PANAS in this study differed from our previous study, however, by having participants respond to the PANAS questions using a slider, rather than a five point scale. This means that only two terms were put at the extreme ends of the slider (i.e. 'very slightly/not at all' and 'extremely' were presented but not 'a little', 'moderately' or 'quite a bit'). The change to use a slider was because van der Sluis and Mellish (2008) observed partipants only using a small part of the possible scale for answers, and within this the five point scale might have lost useful information.

Although our particular experiment focussed on positive affect, we included the negative affect terms partly so that we could detect outliers in our participant set – people who were perhaps extremely nervous about the test or sensitive about their IQ. In fact, we did not find any such outliers.

5 Experiment to Measure Emotional Effects of Positive Feedback

5.1 Set Up of the Study

As stated above, the texts that we presented to our participants were portrayed as giving feedback on an IQ test that the participants had just taken. The IQ test was set up as a web experiment in which participants could linearly traverse through the various phases of the test. An outline of the set up is given in Figure 2. In the general introduction to the experiment, participants were told that the experiment was 'an assessment of a new kind of intelligence test which combines a number of well-established methods that are used as indicators of human brain power'. To make it more difficult for the participant to keep track of how well/poorly she performed over the course of the test, it also said that the test consisted of open and multiple choice questions that had different weight factors in the calculation of the overall score and that would assess various aspects of their intelligence. Subsequently, the participant was asked to tick a consent form to participate in the study. Then a questionnaire followed in which the participant was asked about her age, gender and the quality of her English. She was also asked if she had any experience with IQ tests and how she expected to score on this one. These questions were interleaved with an emotion assessment test (reduced PANAS) in which the participant was asked 'how do you feel right now?'.

After filling out the questionnaire, the participant could start the "IQ test" whenever she was ready. The "IQ test" consisted of 30 questions which she had to answer one at a time. The participant could not skip a question and also had to indicate for each of the questions how confident she was about her answer. The questions that were used for the test were carefully collected from the internet and included items from various tests and games. Different types of questions were used: questions about logical truths, mathematical questions that required some calculations, questions about words and letter sequences, questions including pictures and questions about the participant's personality. They were ordered randomly (but with the same order for each participant).

When the participant had finished the test, she was asked to wait patiently while the system calculated the test scores. When enough calculation time had passed the participant was presented with the test feedback (one of the two texts, regardless of their actual performance). This feedback first explained the test and its type of scoring:

> The Baumgartner test which you have just undertaken tests various kinds of intelligence, for instance, your visual intelligence, your logical-mathematical intelligence and your spatial intelligence. These various aspects of your intelligence contribute to an overall Baumgartner Score. The Baumgartner Score rates your intelligence on a 10-point scale with 10 as the highest possible score. Note that your Baumgartner Score can change over time dependent on experience and practice. Below your test score is presented in comparison with the average score in your age group.

The introduction to the test was followed by either the positively (+1..+7, Figure 1) or the relatively neutrally (01..07, Figure 1) phrased test results. After the participant had processed the feedback, she was asked to fill out one more questionnaire to assess her emotions (i.e. 'How do you feel right now knowing your scores on the test'). This time the simplified PANAS test was interleaved with questions about the participant's results, (e.g. were they as expected and how did she value them), the test (e.g. was it difficult, doable or easy?) and space for comments on the test and the experiment. Finally the participant was debriefed about the experiment and about the goal of the study.

1. General introduction to the experiment;
2. Consent form;
3. Questionnaire on participant's background and familiarity with IQ-test interleaved with a PANAS test to assess the participant's current emotional state;
4. Message: 'Please press the next button at the bottom of this page whenever you are ready to start the intelligence test';
5. IQ test questions;
6. Message: Please be patient while your answers are being processed and your test score is computed. After the result page, you will be asked another set of questions about the test, your performance and the way you feel about it. This information is very important for this study, so please answer the questions as honestly as possible.';
7. Feedback + or 0;
8. Questionnaire: PANAS test to assess how the participants felt after reading the test feedback interleaved with questions about the test, their expectations and space for comments;
9. Debriefing which informed participants about the study's purpose and stated that the IQ test was not real and that their test results did not contain any truth.

Figure 2: Phases in the experiment set up

5.2 Pilot Experiment

A pilot of the experiment was carried out by asking a number of people to try the experiment via the web interface. The main outcomes of this study, in which 11 colleagues participated, was that the experiment was too long. Accordingly, the questionnaires before and after the IQ test (phase 3 and 8 in Figure 2) were shortened. Also the IQ test itself was shortened from 40 to 30 questions.

5.3 Main Experiment: participants and experimental setting

30 participants, all female university students, took the IQ test. All participants except two were in age band 18-24. The exceptions were in age band 25-29 (group +) and 30-34 (group 0). The participants were randomly distributed over group + and group 0 and (for ethical reasons) did the test one by one in a one-person experiment room while the experimenter was waiting outside the room. As soon as the participant indicated that she had finished the task (i.e. stepped out of the experiment room), she was debriefed about the study by the experimenter and was paid with a voucher worth 5 pounds.

5.4 Hypotheses

Since the message of the feedback texts was relatively positive and there is no necessary correlation between positive and negative PANAS scores (Watson and Clark, 1999), we expected the main effects of the texts to be on the average evaluation of the positive PANAS terms. In order to cater for the fact that individuals might differ in their initial positive PANAS scores, we decided to look at the difference of the scores (score after minus score before). Therefore the hypothesis for this study was that participants who received the positively phrased feedback would show a larger change in their positive emotions than the participants who received the neutrally phrased feedback.

	0-group	*+-group*
Negative PANAS terms Before	1.60(.76)	1.58(.68)
Negative PANAS terms After	1.57(.68)	1.31(.45)
Positive PANAS terms Before	3.25(.78)	3.32(.55)
Positive PANAS terms After	3.13(.58)	3.75(.55)

Table 1: Means and Standard deviations (between brackets) for the negative and positive PANAS terms as indicated before and after the IQ test undertaken by participants that received neutral and participants that received positive feedback on their performance.

5.5 Results

Table 1 indicates that on average after they had received their test results, participants in the +-group were more positively tuned than participants in the 0-group. Participants in the +-group also rated the positive emotion terms higher than they had done before they undertook the IQ test. No such results were found for the 0-group. In contrast, compared to their responses before the IQ test, participants in the 0-group rated the positive terms slightly lower after they had processed their neutrally phrased feedback. With respect to the negative PANAS terms, participants in the +-group report slightly less negative emotions after they read their test scores, but none of the differences found in the negative PANAS scores were significant.

A 2 (feedback type) * 2 (before/after) * 2 (positive/negative mean) repeated measures ANOVA was carried out on the average PANAS scores. This showed no main effect of feedback type (+ vs 0) and no main effect of before/after on average PANAS scores. However, there was a highly significant interaction between feedback type and before/after, which indicates that the change in PANAS mean before and after the text was strongly dependent on feedback type[3] ($F(1, 28) = 10.246$, $p < .003$). We interpret this to mean that the (after minus before) value is significantly

[3] An ANOVA test on the positive means only produces a similar result.

	0-group	*+-group*
Alert Before	3.96(.80)	3.17(.99)
Alert After	3.45(.76)	3.65(.75)
Determined Before	3.49(1.02)	3.60(.50)
Determined After	3.50(1.13)	3.74(.61)
Enthusiastic Before	3.52(1.05	3.49(.72)
Enthusiastic After	2.97(.81)	3.84(.66)
Excited Before	2.74(.97)	3.28(.61)
Excited After	2.64(.75)	3.69(.83)
Inspired Before	2.56(1.21)	3.06(.77)
Inspired After	3.06(1.05)	3.81(.78)

Table 2: Means and Standard deviations (between brackets) for the positive PANAS terms as indicated after the IQ test undertaken by participants that received positive and participants that received neutral feedback on their performance.

		0-group	*+-group*
ER			
	not disclosed	1	0
	not so good	0	1
	ok	9	4
	well	4	10
	extremely well	1	0

Table 3: Participant responses when questioned about the results they expected (*ER*) .

greater for the +-group. A two-tailed, two sample t-test verifies this (t = 3.2, $p < 0.004$). We did some post-hoc investigation in an attempt to understand the main result more fully. When looking at the positive PANAS scores in more detail (see Table 2), it turns out that only three of the five positive PANAS terms included in the simplified PANAS test render promising results. Interactions were found for the terms 'alert' (F(1, 28) = 10.291, $p < .003$) and 'enthusiastic' (F(1, 28) = 5.651, $p < .025$). No interactions were found for the terms 'determined' and 'inspired'. For 'inspired' however, we found a main effect of feedback type : (F(1, 28) = 8.755, $p < .006$), which indicates that participants in the +-group could have been more inspired because of their test scores than participants in the 0-group. Not all of these results would be significant if Bonferroni corrections were made.

5.6 The Role of Expectations

It is possible that this result could have been caused by other (systematic but unanticipated) differences between the two groups. In particular, perhaps the result could be caused by a difference in how well the two groups of participants *expected to* perform. As it happens, participants were asked: 'How do you expect to score on an intelligence test?' before they did the test. The answers to this question are summarised in Table 3. This data suggests that participants in the +-group initially had higher expectations. It is difficult to get a consensus from the psychological literature about how this might have affected the results. On the one hand, some studies have shown that positive expectations can have an accelerating effect on a person's actual positive emotional experience (Wilson et al., 2003; Wilson and Klaaren, 1992). Such results might suggest an alternative explanation of the fact that the +-group showed a greater change in positive emotions. On the other hand, it might be argued that subjects with lower expectations would be more surprised (since both texts presented good results) and so their emotions would have been influenced more significantly. That is, if a subject already expects to do well then one would not expect that finding that they actually did well would cause much of a change in their emotions. This would predict that it should be the 0-group that shows the greatest emotion change. Overall, it is hard to know whether the data about expectations should affect our confidence in the experiment result, though it would be worthwhile controlling for initial expectations in further experiments of this kind.

6 Discussion and Future Directions

6.1 Discussion

Compared with the previous study of van der Sluis and Mellish (2008), we expected participants to indicate stronger emotional effects, because the text participants were asked to read was about their own capabilities instead of about something in the world around them which they could think would not affect them. Indeed, this seems to have been the case. In van der Sluis and Mellish (2008), all responses used the lower half of the scale, whereas with the slider our participants indicated values up to both extremes of the range available. Unfortunately, the fact that one set of values is discrete and the other continuous means that it is hard to carry out a simple statistical comparison.

6.2 Future Work

In the study described in the paper, a number of different techniques (e.g. emphasis, vague adjectives and adverbs) were used to phrase the various propositions in the feedback. In future work we aim to identify the relative importance of the individual techniques.

6.3 Conclusion

The fact that we have been able to show a significant difference in the emotions induced by the two texts is very encouraging. It suggests that there *is* a possible methodology for directly evaluating affective NLG and that the tactical concerns with which much of NLG research is occupied are relevant to affective NLG. A similar methodology could perhaps now be used to determine the effectiveness of specific NLG methods and mechanisms in terms of inducing emotions. Although we have now shown that NLG tactical decisions can affect emotions, it remains to be seen what kind of changes in strategy, learning, motivation, etc., can be induced by positive affect and thus how these framing decisions would best be made by an NLG system.

Acknowledgments

This work was supported by the EPSRC platfrom grant 'Affecting people with natural language' (EP/E011764/1) and also in part by Science Foundation Ireland under a CSET grant (NGL/CSET). We would like to thank the people who contributed to this study, most notably Judith Masthoff, Albert Gatt and Kees van Deemter and Nikiforos Karamanis.

References

R. Brown and E. Pinel. 2003. Stigma on my mind: Individual differences in the experience of stereotype threat. *Journal of Experimental Social Psychology*, 39:626–633.

J. Cacioppo, G. Bernston, J. Larson, K. Poehlmann, and T. Ito. 2000. The psychophysiology of emotion. In M. Lewis and J. Haviland-Jones, editors, *Handbook of Emotions*, pages 173–191. New York: Guilford Press.

M. Cadinu, A. Maass, A. Rosabianca, and J. Kiesner. 2005. Why do women underperform under stereotype threat? *Psychological Science*, 16(7):572–578.

D. Graff. 2000. Shifting sands: An interest-relative theory of vagueness. *Philosophical Topics*, 20:45–81.

C. Kennedy. 2007. Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, 30:1–45.

P. Lang. 1980. Behavioral treatment and bio-behavioral assessment: Computer applications. In J. Sidowske, J. Johnson, and T. Williams, editors, *Technology in Mental Health Care Delivery Systems*, pages 119–137. Norwood, NJ: Ablex.

R. Lazarus, A. Kanner, and S. Folkman. 1980. Emotions: A cognitive-phenomenological analysis. In R. Plutchik and H. Kellerman, editors, *Emotion, theory, research, and experience*. New York: Academic Press.

I. Levin, S. Schneider, and G. Gaeth. 1998. All frames are not created equal: A typology and critical analysis of framing effects. *Organizational behaviour and human decision processes*, 76(2):149–188.

A. Mackinnon, A. Jorm, H. Christensen, A. Korten, P. Jacomb, and B. Rodgers. 1999. A short form of the positive and negative affect schedule: evaluation of factorial validity and invariance across demographic variables in a community sample. *Personality and Individual Differences*, 27(3):405–416.

F. Mairesse and M. Walker. 2008. Trainable generation of big-five personality styles through data-driven parameter estimation. In *Proc. of the 46th Annual Meeting of the ACL*.

J. Moore, K. Porayska-Pomsta, S. Varges, and C. Zinn. 2004. Generating tutorial feedback with affect. In *Proceedings of the 7th International Florida Artificial Intelligence Research Symposium Conference (FLAIRS)*.

L. Moxey and A. Sanford. 2000. Communicating quantities: A review of psycholinguistic evidence of how expressions determine perspectives. *Applied Cognitive Psychology*, 14(3):237–255.

L. O'Hara and R. Sternberg. 2001. It doesn't hurt to ask: Effects of instructions to be creative, practical, or analytical on essay-writing performance and their interaction with students' thinking styles. *Creativity Research Journal*, 13(2):197–210.

F. De Rosis and F. Grasso. 2000. Affective natural language generation. In A. Paiva, editor, *Affective Interactions*. Springer LNAI 1814.

F. De Rosis, F. Grasso, and D. Berry. 1999. Refining instructional text generation after evaluation. *Artificial Intelligence in Medicine*, 17(1):1–36.

J. Russell, A. Weiss, and G. Mendelsohn. 1989. Affect grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, 57:493–502.

K. Teigen and W. Brun. 2003. Verbal probabilities: A question of frame. *Journal of Behavioral Decision Making*, 16:53–72.

H. Thompson. 1977. Strategy and tactics: A model for language production. In *Proceedings of the Chicago Linguistics Society*, Chicago.

I. van der Sluis and C. Mellish. 2008. Using tactical NLG to induce affective states: Empirical investigations. In *Proceedings of the fifth international natural language generation conference*, pages 68–76.

D. Watson and L. Clark, 1999. *Manual for the Positive and Negative Affect Schedule - Expanded Form*. The University of Iowa.

D. Watson, L. Clark, and A. Tellegen. 1988. Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54(1063-1070).

T. Wilson and K. Klaaren. 1992. The role of affective expectations in affective experience. In M. Clark, editor, *Review of Personality and Social Psychology*, volume 14: Emotion and Social Behaviour, pages 1–31. Newbury Park, CA: Sage.

T. Wilson, D Gilbert, and D. Centerbar. 2003. Making sense: The causes of emotional evanescence. In I. Brocas and J. Carrillo, editors, *The Psychology of Economic Decisions*, volume 1: Rationality and Well Being, pages 209–233. New York: Oxford University Press.